# An LLM-motivated theory of language

**Ágoston Tóth**
Department of English Linguistics
University of Debrecen
Hungary
toth.agoston@arts.unideb.hu

## Abstract

This work aims to introduce the principles of a linguistic theory motivated by the capabilities of Large Language Models (LLMs). A review of the most relevant literature is included. The core principles of the proposed linguistic theory are identified as 1) an assumed ability to acquire token distribution information, 2) an assumed capability to contextualize distributional information associated with the tokens of processing, 3) the goal is set as the prediction of a token (such as the next token of discourse), while 4) carrying out holistic processing. The proposal utilizes LLMs in the role of modeling tools within the proposed theoretical framework.

## 1 Introduction

The linguistic capabilities of decoder-equipped Large Language Models (LLMs) surprised the general public and the scientific community. Well-trained LLMs generate structurally, semantically, pragmatically sound token sequences in response to user input without prior, explicit linguistic training. Instead, artificial neural networks are trained on unannotated corpora to carry out a distribution-prediction task via contextualizing distributional information.

The linguistic literature has not remained mute on the gap between the new, LLM-derived insights on language processing and existing linguistic theory. The stakes are high for linguists, as we need to identify and understand explanations of how language works, including alternatives. The second section of this paper gives a brief review of this discussion in the literature.

The third section contains my suggestions for the principles that we may use for introducing a linguistic theory that uses LLMs as models of language. I will also identify some pitfalls that seem to decrease the plausibility of current models, which were not originally created for linguistic modeling or theorizing, but for carrying out certain Natural Language Processing tasks.

## 2 The literature on the linguistic status of neural language modeling

Dupre (2021) argues against the idea that machine learning of linguistic patterns could be considered as a contribution to linguistic theories. He underlines the importance of the well-known distinction between linguistic performance (the observable linguistic behavior) and linguistic competence (the underlying system of categories and rules), presenting it in the context of neural language modeling. In the literature, innate capabilities are assumed to exist that help children acquire language despite getting inadequate (performance) data.

Baroni (2022) argues that deep networks (multi-layer artificial neural networks) should be treated as linguistic theories that make ”explicit predictions about the acceptability of linguistic utterances” (ibid.). He also addresses the widespread ’blank slate’ view on deep nets and its relation to innateness theory. He probes network behavior in a long-distance agreement task. Baroni adds that “the appropriate level to represent linguistic knowledge is not algebraic, but massively distributed” (Baroni, 2022).

Bever et al. (2023) argue that deep networks are “poor models of mind and brain for language organization”. Their criticism includes

- the poverty of the stimulus argument: children are exposed to less input than what is needed to acquire complex issues and less than what networks are exposed to,

- the "impossible languages" argument: networks are capable of learning languages that do not exist,
- real syntactic rules and "impossible" ones in unknown languages are handled differently by the brain but not in artificial networks, according to them,
- networks struggle with capturing infinite combinatorial properties of language.

Piantadosi (2024) takes the opposite side. He challenges generative grammar (the title is 'Modern language models refute Chomsky's approach to language'), and uses the spectacular linguistic capabilities of LLMs as counterarguments against core principles of mainstream generative grammar.

- He argues for the viability of the "blank slate" approach: modern language models demonstrate that language acquisition can happen without an innate component, and language learning can occur in unconstrained spaces.
- Prediction, probability: while Chomsky argued that "the notion of 'probability of a sentence' is an entirely useless one" (Chomsky, 1969), LLMs are trained to compute probability in a useful way. This approach allows them to develop gradient, continuous representations.
- LLMs implement the integration of syntax and semantics, which directly challenges the "autonomy of syntax" principle of mainstream generative grammar. The contested principle is based on Chomsky's original claim that syntax alone, as a separate entity, explains why the sentence "Colorless green ideas sleep furiously" is grammatical (Chomsky, 1957).
- LLMs are complex – rather than minimal – representations: Chomsky's approach is to seek a minimal core of representation (cf. "merge"), while LLMs are optimized for utilizing billions of parameters.

Piantadosi (2024) also criticizes mainstream generative grammar for the lack of quantified, implemented theories. He cites the literature to point out that LLMs achieve nearly 90% accuracy in some linguistic tests, such as those on clefts, center-embedding, negative polarity items, and filler-gap dependencies. Piantadosi (2024) also argues that LLMs implement scientific theories of language. He rejects the criticism that these models are unconstrained. Instead, he conceptualizes the weight-setting procedure used in LLM training as a form of theory formation through comparison, with the network searching through a space of possible theories to find the one that best explains the observed data.

Piantadosi's work triggered criticism, more than what we can discuss here. Müller (2025), for instance, argues that LLMs are not linguistic theories, since a "theory contains descriptive and explanatory statements about some part of reality". He goes on to argue that a neural network "may reflect grammatical structures and reject impossible ones, but it does not tell us why this is the case", therefore, a neural network cannot be a theory. He differentiates between theories and models; he conceptualizes a model as "an abstract representation of the relevant part of the reality under consideration" (ibid.).

## 3 A suggested theory featuring trained LLMs as scientific models

In the remainder of the paper, I will use the terms *theory* and *model* in the following way.

A scientific THEORY is an explanatory structure in the framework of which we can substantiate propositions about the object(s) of research. It gives researchers the opportunity to formulate research questions, generate and test hypotheses, integrate empirical findings, discuss statements, build a consensus, to peer-review, etc. Theories may rest on certain assumptions that are usually well-established and agreed-upon. Hypotheses can be corroborated through testing, or falsified by contradictory evidence.

A scientific MODEL is a simplified version of reality that we create and utilize to explore and understand concepts, extract predictions, investigate hypothetical scenarios. Models contribute to theory formulation through the testing options they provide, and their analysis can also generate new theoretical insights.

I suggest that LLMs should be conceptualized as modeling tools for a separate, new theory of language based on certain principles.

I do not argue that LLMs have been developed or trained purposefully to model all aspects of human languages. They may miss features of natural language, but the sheer volume of texts they are trained on, as well as the success with which these models generate text day by day warrant a close inspection.

I suggest the following list of theoretical assumptions (1-4 below) as the core principles of an LLM-motivated theory of language processing:

1. ACQUISITION OF TOKEN DISTRIBUTION. In the suggested theory, language processing is made possible by the acquisition of co-occurrence information for the tokens of the lexicon. First-order token co-occurrence information reveals syntagmatic relations, while second-order token co-occurrence (where the relationship of two tokens is not determined by whether they themselves co-occur, but by the overlap of their contexts) has a potential semantic interpretation according to the distributional hypothesis (Lenci, 2008): words that tend to share a more similar context have more similar meanings.

2. TOKEN CONTEXTUALIZATION. The acquisition of token distribution involves the contextualization of the distributional information calculated for all tokens of the context, where the context includes the tokens that have already been seen in (or generated for) the current discourse. This process operationalizes the distributional knowledge to reach a goal (discussed next).

3. The goal of contextualizing the distributional information is THE PREDICTION OF A TOKEN (the next token, potential next tokens, a missed token, etc.) of the discourse. The network has simultaneous access to all elements of the context.

On the modeling side, LLMs are typically Decoder networks from the Transformer architecture (Vaswani et al., 2017) with input-embedding, multi-layer feed-forward topology, attention blocks, etc., and are trained to solve a distribution-prediction task, such as next-token prediction, by calculating a probability distribution over the items of the token vocabulary. The distributional information that the network acquires is distributed over a large artificial neural network comprised of neurons connected by weighted links. These trainable weights ('parameters') are set during pre-training and fine-tuning processes. Activation weights are modified during training to minimize the error of the network, where the error is the discrepancy between the expected output and the actual token that we see or hear. Weights remain fixed during inferencing.

4. HOLISTIC PROCESSING. We hypothesize that processing is holistic, and that all aspects of language/cognition that have an effect on the distribution-prediction task are handled in parallel. We do not hypothesize that different types of linguistic information (lexical, morphological, syntactic, semantic, pragmatic, etc.) are processed separately or in any particular order. We do not hypothesize the autonomy of syntax or that of any other component.

We do not introduce the equivalent of the performance – competence distinction in this theory. At this time, there is no immediate need to assume the presence of innate linguistic abilities, either. If this position on innateness needs to be revised, on the basis of biological data, then the theory and the models should be modified accordingly.

In this theory of language, with LLMs as modeling tools, there is a path toward aligning the theoretical framework and the model hypotheses with biological observations. The suggested falsification criterion is the lack of correlation between neurolinguistic data and theory-dependent findings. On the side of LLMs, we can probe the network and read neural activation levels and/or extract inter-neural connection weights. It is, of course, more difficult to access a similar set of data in biological networks. Note that the model also allows the researcher to track causality (the chain of events that leads to the activation of neurons, also the subsequent steps, resulting in spatial and temporal activation patterns). Using these probing techniques, we can address research questions similar to these:

a) What are the similarities and differences in activation between events involving different tokens in the same context, and the same token in different contexts? How can we conceptualize these similarities and differences?

b) What does it take for a given element of the token vocabulary to appear as the next output in a given context?

c) Which linguistic generalizations, abstractions (categories, rules), relations (lexical, grammatical, etc.) that have been identified in the linguistic literature can we observe? Can we find an alignment

between these categories and rules with biological data?

When we rely on current LLMs as modeling tools, one of the potential pitfalls is the use of linguistically unmotivated token vocabularies assembled to cover large, multilingual corpora and out-of-vocabulary items. From a linguistic perspective, it does not seem reasonable to think that people collect co-occurrence information about "technical", virtually random tokens – although it is instructive to see that they can do that and make use of ('make sense of') tokens that do not correspond to free or bound morphemes. Also, token vocabularies of trained LLMs are fixed: new tokens cannot be easily introduced into the system without having to retrain the network. I consider these typical LLM design features and possibly others, too, as potential weaknesses in the context of modeling human language processing.

This proposal reframes language as the product of contextualizing distributional information in a holistic system that aims to predict a given token (typically the next token) of a discourse. It has a distinctive view on the lexicon: it conceptualizes it as a set of (information) tokens storing distributional information in a distributed structure, in a neural network, that can operationalize this knowledge based on actual context. It highlights the role of the processing goal, token prediction, and it places all cognitive processes that contribute to the appearance of tokens at the same level in a unified system.

## References


Baroni, M. (2022). On the proper role of linguistically-oriented deep net analysis in linguistic theorizing [Preprint available at arXiv:2106.08694]. In S. Lappin (Ed.), *Algebraic systems and the representation of linguistic knowledge*. Taylor Francis. https://doi.org/10.48550/arXiv.2106.08694

Bever, T. G., Chomsky, N., Fong, S., & Piattelli-Palmarini, M. (2023). Even deeper problems with neural network models of language. *Behavioral and Brain Sciences*, *46*, e387. https://doi.org/10.1017/S0140525X23001619

Chomsky, N. (1957). *Syntactic structures*. Mouton.

Chomsky, N. (1969). Quine's empirical assumptions. *Synthese*, *19*(1), 53–68.

Dupre, G. (2021). (what) can deep learning contribute to theoretical linguistics? *Minds and Machines*, *31*(4), 617–635. https://doi.org/10.1007/s11023-021-09571-w

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Rivista di Linguistica*, *20*(1), 1–31.

Müller, S. (2025). Large language models: The best linguistic theory, a wrong linguistic theory, or no theory at all? *Zeitschrift für Sprachwissenschaft*, *44*(1). https://doi.org/10.18148/zs/2025-2001

Piantadosi, S. T. (2024). Modern language models refute chomsky's approach to language. In E. Gibson & M. Poliak (Eds.), *From fieldwork to linguistic theory: A tribute to dan everett* (pp. 359–384). Language Science Press. https://doi.org/10.5281/zenodo.11351540

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008.